Investigating Student Dropout Risk in Higher Education through Machine Learning

Pedro Mian Parra[1], **Muriel de Souza Godoi**[1], **Francisco Carlos Monteiro Souza**[2], **Alinne C. Correa Souza**[2], **Edgar de Souza Vismara**[2], **Rafael Gomes Mantovani**[1]

[1]Federal University of Technology – Paraná (UTFPR), Apucarana, Paraná, Brazil,

[2]Federal University of Technology – Paraná (UTFPR), Dois Vizinhos, Paraná, Brazil

pedroparra@alunos.utfpr.edu.br,
{muriel,franciscosouza,alinnesouza,edgarvismara,rafaelmantovani}@utfpr.edu.br

Abstract. *In recent years, there has been a significant increase in the student dropout rate in Higher Education. Various reasons are cited, such as difficulty learning the content, the proposed course structure, and lack of financial resources. This study explores machine learning (ML) in the student dropout problem. The experiments were conducted with a dataset from the Academic System of a Federal University listing different academic features of the students. They showed promising results, with Random Forest accurately predicting the student situation with an average F-Score of* 0.959*. However, most relevant features are expected and do not provide any new insight regarding the dropout imminence. Future experiments can fix it with a more robust feature engineering process.*

1. Introduction

Student Dropout is the abrupt interruption of a student's educational journey before completing his/her course [Baggi and Lopes 2011]. It has been a persistent problem in several educational institutions and university courses, resulting in challenges and interfering with university management. It can be understood as a social problem due to its various consequences and ramifications. Among these, the loss of financial resources for public institutions means they might not obtain the return on public investments made, as well as the reduction of revenue in the private sector. In the long term, students who do not receive a diploma may not be able to enter the job market and, consequently, may not contribute to regional development [Evangelista 2017].

Parallel to this, Machine Learning (ML) has been successfully explored in industries, healthcare, and other fields [Mitchell 1997]. Its ability to analyze large volumes of data has generated enormous benefits for those who use them. In education, ML can identify patterns and predict student behavior [Vossen et al. 2023]. The use of ML algorithms in the educational context originated a research area called Educational Data Mining (EDM), where these algorithms and techniques are used to discover patterns and trends in educational data [Kabathova and Drlik 2021]. Through EDM, it is possible to uncover which behaviors and decisions contribute to student success, identify those who are at risk of dropping out or underperforming, personalize content and instruction to meet the specific needs of each student and improve the use of educational resources more efficiently [Ramos et al. 2018].

Thus, this study aims to apply and evaluate the use of ML algorithms to identify students at risk of dropping out of higher education. Data used in experiments were

DOI: 10.5753/sbie.2024.244900

provided by the Academic System of the Federal University of Technology – Paraná (UTFPR), campus of Dois Vizinhos. The research hypothesis investigated states that the induced models can accurately classify a student's situation, labeling them as a dropout or not. The obtained predictions may support managers, coordinators, and directors in creating student policies that minimize this problem.

This paper is organized as follows: Section 2 presents the necessary concepts related to Student Dropout and ML. Section 3 provides the literature review. Section 4 details the experimental methodology employed in experiments with ML. Results are presented and discussed in Section 5. Finally, Section 6 offers conclusions and suggestions for future work.

2. Related Works

In recent years, several studies have investigated the use of algorithms to identify students at risk of dropping out of higher education. [Viana et al. 2022] classified students into Dropped Out and Graduated categories, generating models trained for each semester window. The experiments used data from the Computer Science and Information Systems courses at the Federal University of Piauí (UFPI). The algorithms used are Random Forest (RF), Decision Tree (DT), Extra Trees (ET), Multilayer Perceptron (MLP), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Gaussian Naive Bayes (GNB). The results achieved an accuracy exceeding 86% in cross-validation, reaching 95.7% in the 5th period. Different algorithms perform best at various periods, with RF excelling in the 1st and 2nd periods, while ET and MLP are optimal for the 4th and 6th periods.

In his master's thesis, [dos Santos 2022] investigated student dropout at the Federal University of São Carlos (UFSCar). The author modeled the problem as a binary classification task and used data from academic systems and online questionnaires. Results indicated vital characteristics and the context of dropout in the institution, enabling identifying students at risk of dropping out in the first years of the course. The best algorithm was *LightGBM* with an accuracy of 0.83 and *F1-score* of 0.80, respectively.

The study of [Marques et al. 2020] unravels the causes of academic dropout in the Computer Science course at UFERSA. The authors use the K-Means algorithm and a clustering technique over a database extracted from the Integrated Management System for academic activities. The results indicate that students who tend to drop the course have brown skin color, have many failures during the course, and have a low monthly income.

[FIlho et al. 2020] addressed student dropout on the most affected campuses of the Instituto Federal do Ceará, where rates were higher than 40%. Handling it as a binary classification problem, they used data from five campuses obtained through the *IFCE em Números* platform from 2015 to 2019. The data were preprocessed, and four ML algorithms were evaluated. The best performance was observed with *Gradient Boosting*, reaching *F1-Score* of 0.87 on the Quixadá campus. Similarly, [de Almeida Teodoro and Kappel 2020] investigated the probability of dropout among students at public higher education institutions in Brazil. Using data from INEP, five ML algorithms were applied, with RF standing out as the most effective, with an accuracy rate of around 80%. The results revealed that characteristics such as age, participation in extracurricular activities, and course workload determine dropout factors.

Finally, the study by [Martins et al. 2023], carried out at the Federal University of Juiz de Fora (UFJF), investigates student dropout in higher education, focusing mainly on late dropout, where the student drops out after the fifth period. Using data from in-person undergraduate students between 2003 and 2020, it addressed different perspectives on student dropout, including its internal and external causes and the resulting economic and social impacts. The researchers employed algorithms such as DT, K-Nearest Neighbors (KNN), LR, and RF based on robust data preprocessing, transformation, and selection methodologies. RF was the most effective in predicting late dropout in the general and finalist datasets, with F1 scores of 0.93 and 0.88, respectively.

3. Experimental Methodology

Figure 1 presents the experimental methodology adopted in the experiments. The figure represents the complete *pipeline*[1] and all the sub-tasks necessary for the elaboration of the ML solution, from data acquisition to the model evaluation. Each of the following subsections gives details of these sub-tasks.

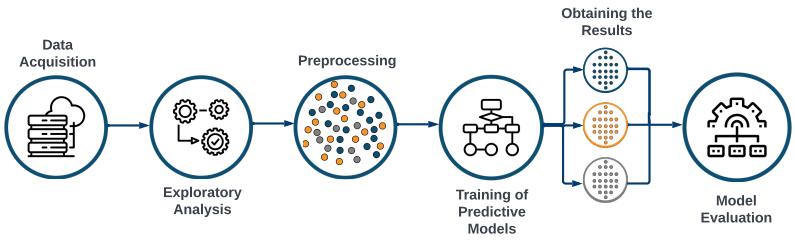

Figure 1. ML pipeline employed in the experiments.

3.1. Dataset

Data were extracted from the Academic System of the Academic System of the Federal University of Technology – Paraná (UTFPR), campus of Dois Vizinhos, approved by its Research Ethics Committee, focusing on information such as academic performance, performance coefficient, and sociodemographic data of students. There are seven active undergraduate courses on campus the with active records. All the identifiers and sensitive information have been removed, and an anonymized version of the data was used in experiments. All the features extracted from the Academic System are listed in Table 1.

For this study, the terminology used to label the student situations will follow the standard present in the UTFPR's Academic System, where:

- **Dropout:** a student who canceled their enrollment;
- **Graduated:** a student who completed their degree and received their diploma;
- **Regular:** a student who is regularly enrolled and progressing through their degree;
- **Suspended:** a student who temporarily suspended their enrollment;
- **Transferred:** a student who transferred to another course and/or campus.

The university keeps semester records of student performance. However, only the latest records were kept in the dataset. For example, a student in the 8th semester of a course has

[1] A *Pipeline* is a sequence of interconnected data processing and modeling steps, developed to automate, standardize, and accelerate the process of creating, training, evaluating, and implementing models.

Table 1. Features obtained from the Academic System and included in the datasets of this study

Analyzed features	
Level of Education	Degree
Frequency	Course
Shift	Gender
Year of Admission	Year/Semester of Status
Freshman	Stricto Sensu Category
Absolute Academic Performance	Normalized Academic Performance
Course Code	INEP Course Code
Graduation Date	Conclusion Date
Admission Method	Age
Changed Course (Same Campus)	Changed Course (Different Campus)
ENEM Humanities Score	ENEM Language Score
ENEM Math Score	ENEM Science Score
ENEM Writing Score	Final SISU Score
Number of Entries in Other Undergraduate Courses	Number of Entries in Same Course
Number of Approved Subjects	Number of Failed Subjects by Attendance
Number of Failed Subjects by Grade	Course Admission Order
Country of Birth	Period
Probable Dismissal	Teaching Regime
Partial Retention	Total Retention
Semester of Admission	Target
Quota Type	Total Semesters Completed
Total Course Semesters	

only their 8th semester record included in the dataset. This approach was chosen because the most recent record provides a consolidated view of the student's academic progress and current standing within the course. Using historical records from previous semesters could introduce temporal biases and unnecessary complexity, as the primary focus is on understanding the students' most recent academic status. Figure 2 depicts the amount of records of each type in the Academic System.

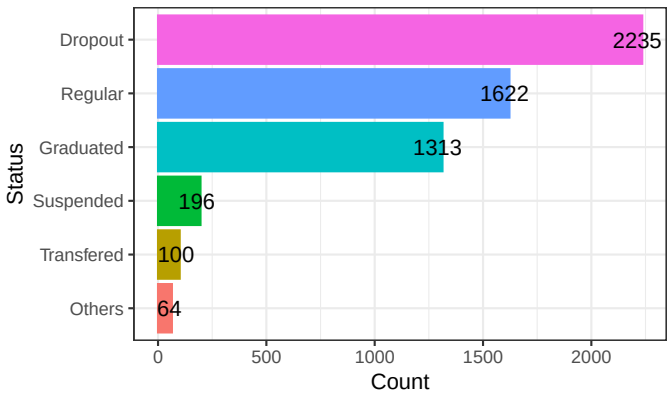

Figure 2. Number of Students in Each Situation

We handled the dropout prediction as a binary classification task and varied what we considered as *Regular* and *Dropout* instances. Five different tasks were created from the original data, all of them listed in Table 2. For each task, the table presents: the task

number, its acronym and correspondent classes, the number of features, its number of examples, how many examples belong to the Regular class, the number of examples from the Dropout class, and the dropout (imbalancement) rate.

Table 2. Binary classification tasks generated for the experiments

Task	Acronym (Regular vs Dropout)	Features	Examples	Class Regular	Class Dropout	Dropout rate
1	Graduated+Regular vs Dropout	43	5170	2935	2235	0.43
2	Graduated+Regular vs Dropout+Suspended	43	5366	2935	2431	0.45
3	Regular vs Dropout	43	3875	1622	2235	0.58
4	Regular vs Dropout+Suspended	43	4053	1622	2431	0.60
5	Graduated vs Dropout	43	3548	1313	2235	0.63

3.2. Preprocessing

All the identifiers and sensitive information have been removed, and an anonymized version of the data was used in experiments. Missing values were also imputed according to the feature type: numerical values are imputed with the column median value, while categorical values demand for a new category. We also removed nearly constant features through a "constant threshold" cc value, filtering features that do not provide useful variation for model learning. The considered value for cc was 0.05. So if the standard deviation of a dataset's column is lower than this value, this column will be removed. In the last step, we removed highly correlated (redundant) features through the Spearman correlation score and a threshold cr. The considered value was $cr = 0.85$. Then, if two features have an absolute correlation value higher than the threshold value, one of them is removed.

3.3. ML Algorithms

Four ML algorithms were included in the experiments: k-Nearest Neighbors (kNN), Naïve Bayes (NB), Decision Trees (DTs) and Random Forest (RF). These algorithms follow different learning biases and can learn different decision surfaces on data. They can also be considered "white-box" algorithms since we can extract useful information from the induced models and interpret their predictions, which is highly important for decision-making processes. All of them were implemented in Python using the scikit-learn library. We did not perform hyperparameter tuning in the experiments, keeping the algorithms' default hyperparameter values.

3.4. Evaluation and Reprodutibility

Since some of the datasets are imbalanced, we performed a 10-fold stratified Cross-Validation (CV) resampling for model evaluation. The experiments were also repeated with 10 different seeds to better assess the induced models. The evaluation measures selected to assess models were the Balanced Per Class Accuracy (BAC) and F1-Score. We also employed non-parametric statistical tests to compare the performance of the induced models. The Friedman-Nemenyi test (α = 0.05) was applied to compare different techniques on different datasets. When paired comparisons were required, we employed the Wilcoxon test with the same $alpha$ value [Santafe et al. 2015].

4. Results and Discussion

4.1. Overall results

Figure 3 depicts the best F1-Score values obtained during the experiments. In the figure, all the algorithms are listed on the x-axis, while the tasks are listed on the y-axis. The deeper the shade of blue in a cell, the better the induced model. The cell values represent the F1-Score achieved by the algorithm for the corresponding task.

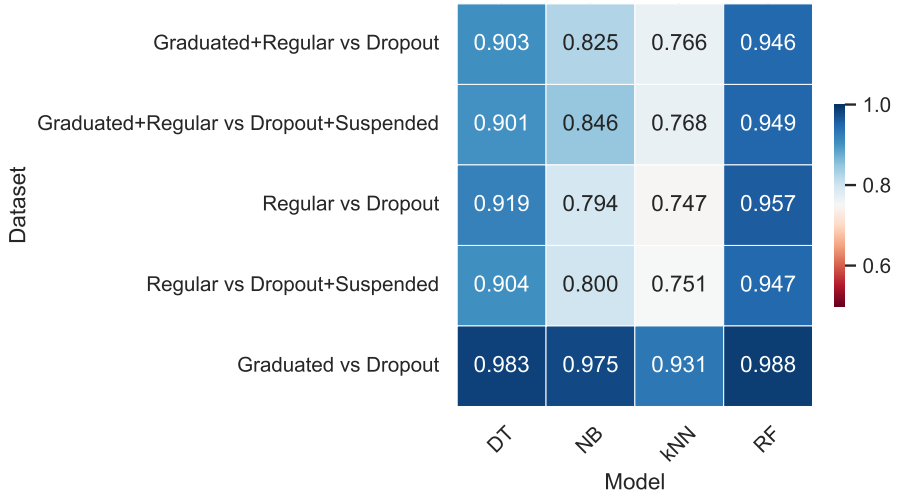

Figure 3. Best F1-Score values obtained by the induced models in the all the tasks.

Figure 4. Comparison of the predictive F1-Score values of the induced models according to the Nemenyi test. Groups of algorithms that are not significantly different (at $\alpha = 0.05$) are connected.

Among the algorithms, kNN performed the worst, with an average F1-Score 0.793. It was followed by NB (0.848) and DT (0.922). The best average results were achieved by RF (0.959), which outperformed all other algorithms in the five tasks. In order to compare them across all tasks, we applied the Friedman test with a significance level at $\alpha = 0.05$. The null hypothesis states that all algorithms induced are equivalent concerning the F1-Score values. If the null hypothesis was rejected, the Nemenyi post-hoc test was applied, stating that the performances of two different techniques are significantly different if the corresponding average ranks differ by at least a Critical Difference (CD) value. Figure 4 shows the obtained CD diagram. Algorithms are connected when there are *no* statistically significant differences between them. In the figure, we can see that all the algorithms present significant differences between each other, with RF outperforming all of them.

In terms of tasks, task 5 is the "easiest" one, with an average F1-Score of 0.983. All the algorithms presented high F1-Score values, but it is also the smallest in terms

of examples (3548), discarding almost two thousand valid samples. The other tasks presented similar results, all of them ranging from $[0.85, 0.87]$. From a management perspective, task 2 (Graduated + Regular vs Dropout + Suspended) is the best fit for a decision support system. It includes all the available examples from the Academic System and considers various scenarios related to student dropout. This task is particularly relevant because it encompasses the most critical and impactful situations within the academic system, making it highly valuable for day-to-day course management. In this task, RF obtained an average F1-Score of 0.959 and is statistically superior to the others, stated by a Wilcoxon paired-test with $\alpha = 0.05$ (95% of significance).

4.2. The best model predictions

Figure 5 presents RF confusion matrix on task 2 (Graduated+Regular vs Dropout+Suspended). There, one may note that both classes presented high accurate rates: 0.965 for Regular and 0.94 for Dropout. Thus, the higher number of examples helps the model to correctly define the decision boundaries. Here, the positive class describes a *Dropout* student. So, a false positive (FP) is a regular student identified as a dropout, while a false negative (FN) is a dropout student identified as a regular one. Checking the figure, most of the misclassifications are FNs (145). From a management perspective, minimizing it is crucial and could be improved in future experiments.

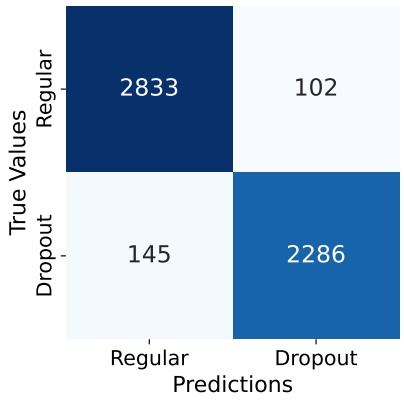

Figure 5. RF confusion matrix on task 2

To verify which features help identify each class in the learning process, the importance values based on the permutation of the RF algorithm were taken as a basis. Figure 6 presents the most relevant features, whose relative importance in terms of Gini index ≥ 0.01. Top-3 features were: the number of approved courses, the absolute performance coefficient, and the total semesters completed. They are expected due to the data nature: regular students tend to stay longer in the courses, are approved in a greater number of subjects, and as a consequence, they have a higher overall coefficient (average grade) than new or dropout students.

Several interesting features can be observed. The Number of Courses that Failed Due to Attendance, Total Retention, and Number of Courses that Failed Due to Grades are significant indicators of academic struggle. Students with a higher number of failures, whether due to grades, attendance, or both, are more likely to drop out of higher education. In contrast, features such as Year of Entry and Year/Semester of Situation are less critical,

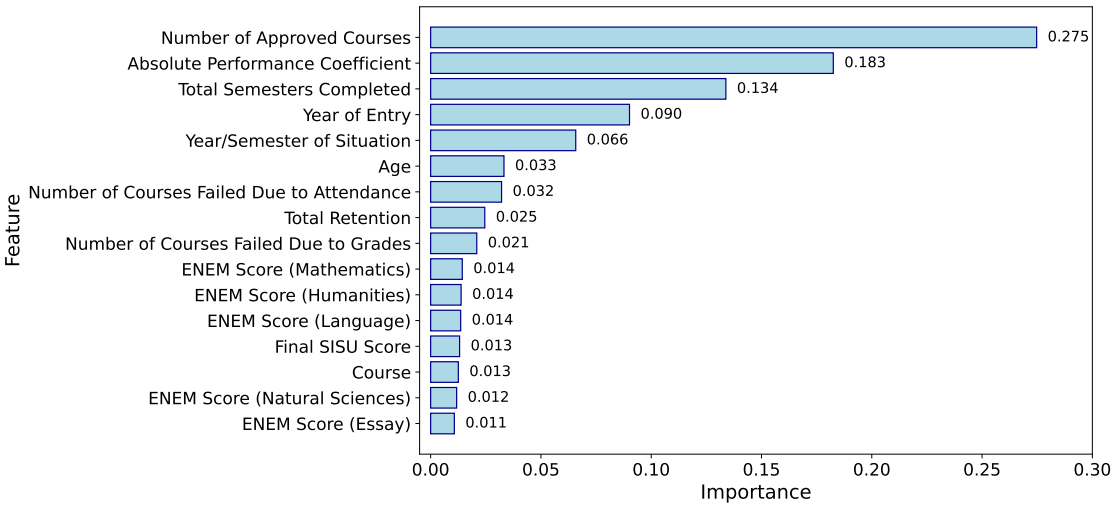

Figure 6. Feature Importance in Random Forest with 0.85 Correlation, 0.05 Consistency Thresholds

primarily reflecting historical enrollment trends. Notably, this university experienced its highest intake of students between 2015 and 2017, which correlates with an increased number of dropouts, as dropout rates are typically higher in the early semesters.

Given that many of the identified "relevant" features do not offer new insights into the dropout situation, there is a need for a deeper understanding of the features influencing dropout rates. Future experiments should employ a more robust dataset with features that are less prone to overfitting, allowing for more accurate predictions and analysis.

5. Conclusion

In this study, we performed ML experiments to detect dropout students in a dataset from the Federal University of Technology – Paraná (UTFPR), campus of Dois Vizinhos. Data was preprocessed and different tasks were generated considering what a dropout student would mean. A total of four "white-box" algorithms were evaluated and the best results were obtained by RF with an average F1-Score of 0.959. However, when analyzing the induced model's most relevant features do not provide any new insight regarding the dropout situation. Most of the rules explore features like the number of approved courses/disciplines, the absolute performance coefficient, and the total semesters completed.

Future research includes several key enhancements and explorations. First, we plan to perform a complete feature engineering process, incorporating socioeconomic features into the dataset, and consider including all student records across multiple semesters. Another potential avenue is to implement a temporal learning strategy, monitoring student dropout probabilities over time. A final step will be the automation of the experimental pipeline, utilizing optimization techniques such as genetic algorithms to efficiently identify the best models without exhaustive search.

References

Baggi, C. A. D. S. and Lopes, D. A. (2011). Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. *Revista da Avaliação da Educação Superior (Campinas)*, 16:355–374.

de Almeida Teodoro, L. and Kappel, M. A. A. (2020). Aplicação de técnicas de aprendizado de máquina para predição de risco de evasão escolar em instituições públicas de ensino superior no brasil. *Revista Brasileira de Informática na Educação*, 28:838–863.

dos Santos, R. S. S. (2022). Evasão escolar universitária e estratégias de intervenções para retenção do estudante: Um estudo de caso na universidade federal de são carlos. Master's thesis, Universidade de São Paulo.

Evangelista, R. W. (2017). *Estudo da evasão do Bacharelado em Humanidades da UFVJM: causas e consequências*. PhD thesis, Programa de Pós-Graduação em Educação, Universidade Federal dos Vales do Jequitinhonha e Mucuri, Diamantina.

FIlho, F. W. B. H., Vinuto, T. S., and Leal, B. C. (2020). Análise de classificadores para predição de evasão dos campi de uma instituição de ensino federal. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 1132–1141, Porto Alegre, RS, Brasil. SBC.

Kabathova, J. and Drlik, M. (2021). Towards predicting student's dropout in university courses using different machine learning techniques. *Applied Sciences*, 11(7).

Marques, L. T., Marques, B. T., Rocha, R. S., e Silva, L. C., Queiroz, P. G. G., and de Castro, A. F. (2020). Evasão acadêmica e suas causas em cursos de bacharelado em ciência da computação: Um estudo de caso na ufersa. In *Anais dos Workshops do IX Congresso Brasileiro de Informática na Educação (CBIE 2020)*, pages 1042 – 1051.

Martins, C. V. M., Lacerda, F. C., do Carmo, I. P., da Silva, E. V. S., Alves, T. O. M., Gomes, J. M., and Campos, R. S. (2023). Modelos de previsão de evasão tardia na graduação de uma universidade pública. pages 41–50. Sociedade Brasileira de Computação - SBC.

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill International Editions. McGraw-Hill, Nova York.

Ramos, J., Silva, J., Prado, L., Gomes, A., and Rodrigues, R. (2018). Um estudo comparativo de classificadores na previsão da evasão de alunos em ead. *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)*, 29(1):1463.

Santafe, G., Inza, I., and Lozano, J. A. (2015). Dealing with the evaluation of supervised classification algorithms. *Artificial Intelligence Review*, 44:467–508.

Viana, F. S., Santana, A. M., and de Andrade Lira Rabêlo, R. (2022). Avaliação de classificadores para predição de evasão no ensino superior utilizando janela semestrais. In *Anais dos Workshops do XI Congresso Brasileiro de Informática na Educação (CBIE 2022)*, pages 908 – 919.

Vossen, L. V., Santos, M. S., Frigo, L., and Gasparini, I. (2023). Dropoutless: plataforma colaborativa de predição de evasão. In *Anais do XVIII Simpósio Brasileiro de Sistemas Colaborativos*, pages 193–201, Porto Alegre, RS, Brasil. SBC.